STRIDE – An Integrated Standards-Based Translational Research Informatics Platform

Henry J. Lowe MD, Todd A. Ferris MD, MS

Penni M. Hernandez ND, RN, Susan C. Weber PhD

Stanford Center for Clinical Informatics, Stanford University, Stanford CA

Abstract

STRIDE (Stanford Translational Research Integrated Database Environment) is a research and development project at Stanford University to create a standards-based informatics platform supporting clinical and translational research. STRIDE consists of three integrated components: a clinical data warehouse, based on the HL7 Reference Information Model (RIM), containing clinical information on over 1.3 million pediatric and adult patients cared for at Stanford University Medical Center since 1995; an application development framework for building research data management applications on the STRIDE platform and a biospecimen data management system. STRIDE's semantic model uses standardized terminologies, such as SNOMED, RxNorm, ICD and CPT, to represent important biomedical concepts and their relationships. The system is in daily use at Stanford and is an important component of Stanford University's CTSA (Clinical and Translational Science Award) Informatics Program.

Introduction

Informatics has emerged as an important contributor to effective clinical and translational research (CTR), particularly within NIH's Clinical and Translational Science Awards (CTSA) program[1]. In 2003 the Institute of Medicine's (IOM) Clinical Research Roundtable published a report on challenges facing the national research enterprise[2]. They recommended the development of information systems "designed to meet the needs of clinical research". In 2005, as a follow up to this report, Payne et al conducted a qualitative study of how informatics was being used to support translational research at several leading academic medical centers[3]. They found that fragmentation of resources frequently hindered the effective use of informatics in CTR and that Electronic Health Record (EHR) systems were of limited value in supporting translational research. In 2008 DiLaura et al looked at the use of informatics in the clinical research enterprise at 52 U.S. sites[4] and found little progress, other than in planning, strategy and governance, since a similar study conducted by them in 2005. In 2008, many of the same challenges listed in the IOM report were again identified by Williams et al in their report on lessons learned from the Clinical Research Networks Initiative[5]. A 2009 review on reuse of clinical data for CTR emphasized the need for integrating clinical research projects with data repositories built up during documentation of routine clinical care[6]. Over the past few years a number of integrated sharable CTR informatics models have emerged, including caBIG[7] and i2b2[8].

In 2002, Stanford University School of Medicine launched a new strategic plan called "Translating Discoveries", that included the goal to "promote the development of those interdisciplinary research collaborations most likely to lead to improved health through the exploitation of new knowledge of biosciences and disease mechanisms". To support its broad goals related to CTR, the School created the Stanford Center for Clinical Informatics (SCCI), tasked with developing an enterprise-level informatics architecture supporting the needs of CTR, leveraging on the Stanford electronic health record. In 2003, SCCI began a multiyear research and development project to create an integrated, standards-based informatics platform addressing three major barriers to effective CTR: (1) efficient access to clinical data for research purposes; (2) delivery of robust research data management solutions and (3) availability of an enterprise-level system for managing and discovering biospecimen data. The new system is called STRIDE (Stanford Translational Research Integrated Database Environment). This paper outlines the architecture and functionality of STRIDE, a system that is now in daily use at Stanford University Medical Center (SUMC) and is a core component of Stanford's CTSA informatics model.

Methods

STRIDE is built on the Oracle 11g relational database platform and uses an n-tiered architecture. Data is stored using an Entity-Attribute-Value (EAV) model[9] and is represented by object-oriented data structures (entities, roles and acts) derived from the HL7 Reference Information Model (RIM)[10]. The system includes a Master Person Index (MPI) that is dynamically populated from clinical, research and biospecimen data. The STRIDE MPI is the only fully integrated enterprise MPI operating at SUMC.

The STRIDE physical database layer is organized into three logical database partitions: (1) a Clinical Data Warehouse (CDW); (2) Research Database Management supporting multiple logically separate research databases and (3) the Biospecimen Data Repository, which supports multiple separate biospecimen databases. All three components rely on the same underlying architecture and services. Data can be linked across all three partitions. For example, with IRB approval, research and/or biospecimen data can be merged with clinical data from the Clinical Data Warehouse.

STRIDE's semantic layer consists of a framework supporting multiple terminologies, including ICD9-CM, ICDO, CPT, RxNorm and SNOMED. This mixed terminology model supports standards-based data entry, data integration, hierarchical concept-based retrieval and data interoperability. As an example, using RxNorm to represent pharmacy data in STRIDE allows the system to merge drug information from the two different vendor drug models used at SUMC, with dynamic integration of pediatric and adult medication orders within the CDW. Additionally, the semantic model allows linkage from RxNorm to SNOMED drug classes.

STRIDE applications use a suite of Java Swing clients communicating with a set of services for managing demographic, clinical, research and reference data (including controlled vocabularies). These software services are organized in a Service Oriented Architecture (SOA) with the baseline services serving not only client applications but also the clinical and research data services within the core system stack (Figure 1).

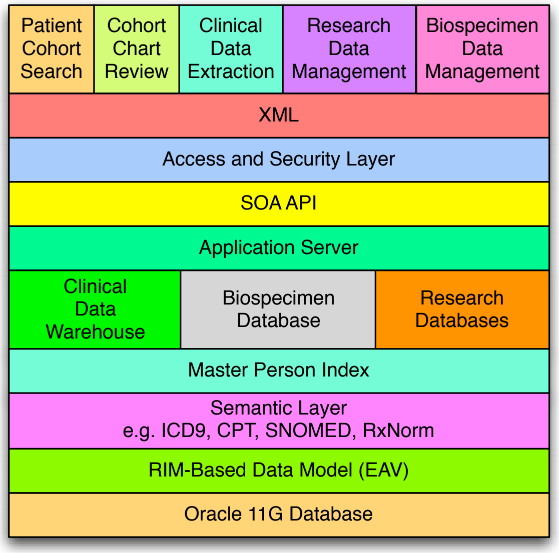


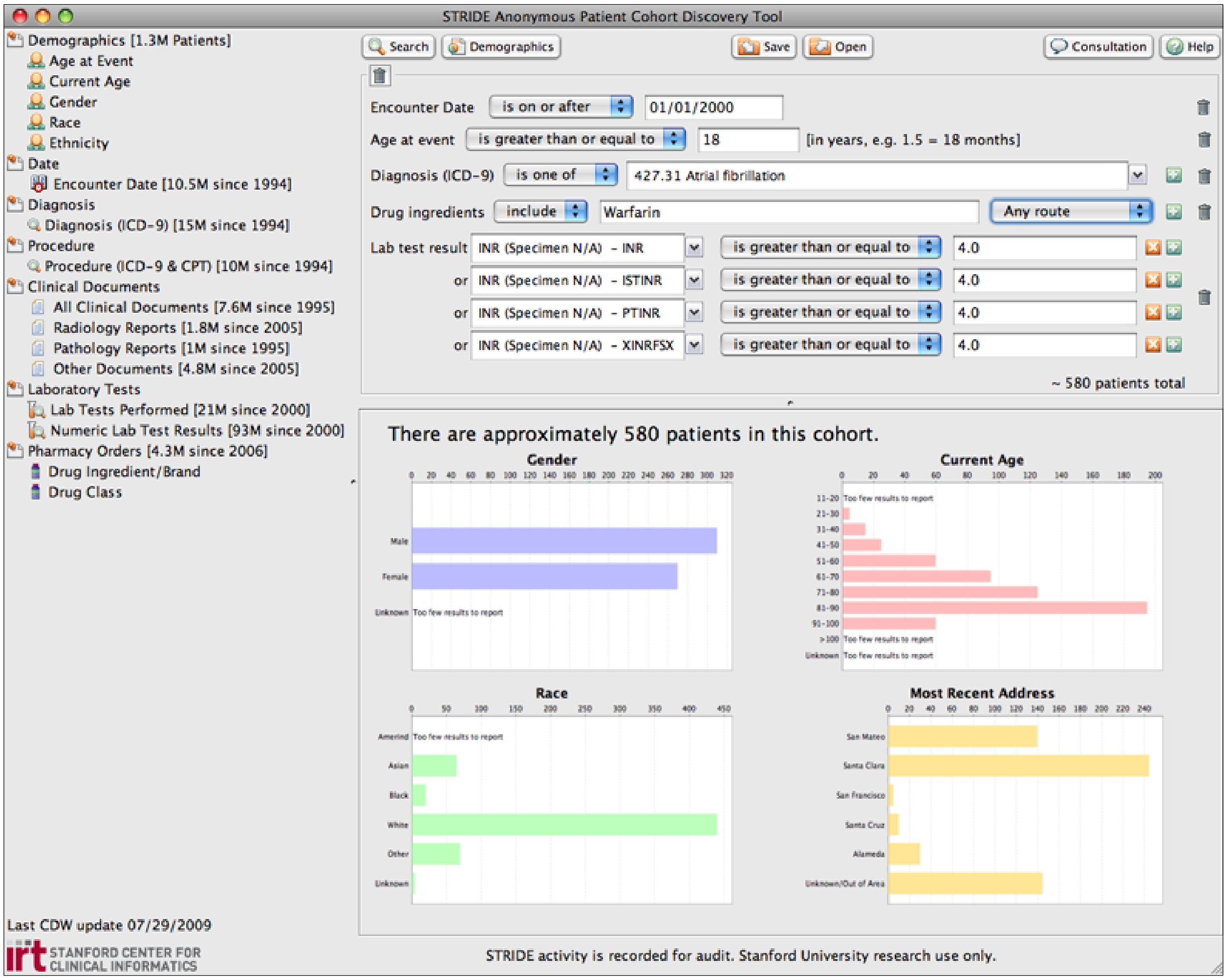
Figure 1. STRIDE Architecture Stack

STRIDE client applications, written in Java, are created using reusable components to facilitate rapid application development. A rich application programming interface (API) permits client applications to dynamically specify the content of data retrieval queries which are then translated on the server side into SQL scoped by the data visibility rules in the STRIDE access and security layer, resulting in dynamically reconfigurable data sets called Virtual Private Databases[11] (VPD) which support highly detailed access control rules.

As of July 2009 the STRIDE CDW contained fully-identified merged demographic and clinical information on over 1.3 million pediatric and adult patients cared for at Stanford University Medical Center (SUMC) since 1995. Table 1 shows the major classes of data stored in the CDW, along with counts and starting year. All clinical documents and reports are full-text indexed and searchable using Oracle Text[12]. The CDW was initially populated with legacy clinical data extracts from the two SUMC EHR systems (Lucile Packard Children's Hospital at Stanford uses Cerner and Stanford Hospital and Clinics uses Epic). The CDW is now populated, in real-time, using HL7 message feeds from a variety of SUMC clinical systems. These version 2.x HL7 messages are parsed by STRIDE to create HL7 RIM-based data structures and linked to patient identifiers in the STRIDE MPI. Data from the Social Security Death Index (SSDI) is integrated into the CDW.

Clinical Encounters	10.5 Million (Since 1994)
Diagnoses (ICD9)	15 Million (Since 1994)
Procedures (CPT, ICD9)	10 Million (Since 1994)
Radiology Reports (Full-Text)	1.8 Million (Since 2005)
Pathology Reports (Full-Text)	1 Million (Since 1995)
Clinical Documents (Full-Text)	4.8 Million (Since 2005)
Laboratory Test Results	93 Million (Since 2000)
Inpatient Pharmacy Orders	4.3 Million (Since 2006)

Table 1. STRIDE CDW Data (July 2009)


STRIDE Anonymous Patient Cohort Discovery Tool
Demographics [1.3M Patients]
Age at Event
Current Age
Gender
Race
Ethnicity
Date
Encounter Date [10.5M since 1994]
Diagnosis
Diagnosis (ICD-9) [15M since 1994]
Procedure
Procedure (ICD-9 & CPT) [10M since 1994]
Clinical Documents
All Clinical Documents [7.6M since 1995]
Radiology Reports [1.8M since 2005]
Pathology Reports [1M since 1995]
Other Documents [4.8M since 2005]
Laboratory Tests
Lab Tests Performed [21M since 2000]
Numeric Lab Test Results [93M since 2000]
Pharmacy Orders [4.3M since 2006]
Drug Ingredient/Brand
Drug Class
Last CDW update 07/29/2009
STANFORD CENTER FOR CLINICAL INFORMATICS
Search
Demographics
Save
Open
Consultation
Help

Encounter Date: is on or after 01/01/2000

Age at event: is greater than or equal to 18 [in years, e.g. 1.5 = 18 months]

Diagnosis (ICD-9): is one of 427.31 Atrial fibrillation

Drug ingredients: include Warfarin Any route

Lab test result: INR (Specimen N/A) - INR is greater than or equal to 4.0

or INR (Specimen N/A) - ISTINR is greater than or equal to 4.0

or INR (Specimen N/A) - PTINR is greater than or equal to 4.0

or INR (Specimen N/A) - XINRFSX is greater than or equal to 4.0

~ 580 patients total

There are approximately 580 patients in this cohort.

STRIDE activity is recorded for audit. Stanford University research use only.

Figure 2. STRIDE Anonymous Patient Cohort Discovery Tool

Stanford researchers can search the CDW to identify potential research patient cohorts using a self-service visual query tool called the Anonymous Patient Cohort Discovery Tool (Figure 2). This Java application, inspired by the earlier work of Murphy et al[13], uses an intuitive drag and drop interface to create and execute queries in real time, answering the question "does the STRIDE CDW contain a cohort of patients with these attributes?" No individual patient data is exposed by this tool and binning[14] is used to prevent identification of individual patients within small cohort sets. All research patient cohort searches are logged for audit purposes. Patient cohorts of interest can be saved within the central STRIDE system for later use in IRB-approved chart review or research data set extraction.

STRIDE also provides a research cohort data review tool designed to allow examination of clinical data to help determine if patients are suitable for inclusion in research studies. Each user can maintain a personal library of research patient cohorts within the system with access to protected health information (PHI) controlled by IRB approval. STRIDE is operated as a Stanford IRB-approved research protocol, with a waiver of informed consent and HIPAA authorization, that permits all clinical data captured as part of routine patient care at SUMC to be transferred to the STRIDE CDW. The transfer of clinical data to the CDW is further governed by a detailed agreement between Stanford's two hospitals and the Stanford University School of Medicine, where the project is based. The STRIDE project works closely with the Stanford University IRB and School of Medicine Privacy Officer to ensure data is only used in an authorized manner. Identified clinical data in the CDW is only released to IRB-approved research studies that have received the appropriate IRB approval. De-identified data is made available for Stanford research projects that has been determined to qualify as a non-human subject research study. Patient contact information is only released to studies that have a detailed IRB approved recruitment strategy. STRIDE includes a robust logging framework and audit capabilities.

Rich research data management applications, incorporating imaging data can be developed on the STRIDE platform. The STRIDE application development workbench offers a wide variety of both stock and custom interface components, ranging from simple lists of values, numeric or text input fields and search-as-you-type data discovery widgets that permit users performing research data entry to rapidly search reference data for an appropriate value. The reference data underlying the data discovery widgets can be patient or physician identifiers or any of the controlled vocabularies used to codify data where appropriate. The orchestration of the underlying services, namely the MPI (Master Person index), tissue and image search, data extraction, data profile discovery, PACS image integration, data retrieval and data persistence services, is achieved by creating XML descriptor files specifying which services a given application requires. For applications using the data retrieval and persistence service, which includes all research data management applications, additional XML descriptors enumerate the labeling, ordering and placement of each reusable data capture component, permitting rapid development of custom data capture applications.

The STRIDE Biospecimen Data Repository provides registration, tracking and management functions for multiple separate biospecimen banks. Individual banks can store and manage biospecimen data privately within STRIDE, while aggregate biospecimen data can be searched across all banks. This allows the creation of a virtual enterprise biospecimen data repository to which data from additional banks can be added incrementally. As of July 2009, STRIDE contained data on approximately 50,600 biospecimens at Stanford (including tissue, DNA, bone marrow and a variety of blood components), managed in multiple banks. As all stored biospecimen data shares the same STRIDE MPI and the same core data elements (with clinical and pathologic data represented using SNOMED), aggregate anonymous searching across all biospecimen banks is supported, allowing Stanford researchers to discover and request biospecimen samples with specific characteristics. With appropriate IRB approval, biospecimen data can be linked to specific clinical data elements in the STRIDE CDW.

Results

The STRIDE project is a multiyear process of incremental development, testing and deployment of functional components to meet Stanford's specific CTR needs. The first research data management applications were launched in mid-2005. The biospecimen component was released in late 2005 and the first production release of the CDW occurred in early 2008. Usage has grown as system functionality has increased. There have been over 300 uses, by about 70 individual users, of the Patient Cohort Discovery Tool since its first release in mid 2008. There are currently approximately 90 users of various STRIDE research data management applications. To most effectively connect researchers to STRIDE services we have developed a free Informatics Consultation service. In many cases consultations result in the delivery of specific STRIDE-related services, such as development of a research data management solution or provision of clinical data for research purposes. These consultations also provide us with valuable input as to what additional services are needed and how existing services might be enhanced. For example, user feedback suggested that in addition to current age (i.e. calculated age at time of query), the age of the patient at the time of a clinical encounter was an additional criterion that would be valuable when defining research cohort queries. This issue, which has been described in other CDWs supporting research[15], was addressed and is now supported by the STRIDE Anonymous Patient Cohort Discovery Tool.

Building a data warehouse is often an iterative process[16] in which the developer attempts to optimize resource allocation, data availability, technical complexity, scalability and end user needs to deliver incremental functionality, while maintaining the institutional support needed for an expensive multi-year initiative. The STRIDE CDW has been managed as a phased project in which, each academic year, we add new clinical data and deploy new system functionality. For example in the current academic year (September 2008-August 2009) we have added SUMC inpatient pharmacy data to the CDW; deployed a new version of the Patient Cohort Discovery Tool that supports pharmacy-related search criteria (e.g. drug ingredients, drug classes and administration routes) and are evaluating a Research Cohort Data Review Tool. We are currently weighing the question of how much of the available SUMC clinical data to aggregate within the CDW versus accessing some data elements, as needed, in situ (i.e. from within the EHR itself). If the sole use of the CDW is providing access to clinical data extracts for research purposes, then a hybrid model may be adequate. However, if use of the CDW extends to tasks such as expanded research cohort identification or clinical data mining a high degree of integration is required.

To date multiple research data management applications and databases have been built on the STRIDE platform, ranging from one large multimedia-capable system supporting an entire academic department; various longitudinal patient registries (some incorporating radiology imaging data) and several less complex systems supporting small to medium sized research projects. In this latter category, we feel that the cost-benefit analysis may not warrant continued use of STRIDE as a solution and we are currently evaluating REDCap[17] as an alternative institutional solution for managing data in small to medium sized research projects.

The STRIDE Biospecimen Data Management system is as much a cultural strategy as it is a technology solution. Our hope is to encourage the owners of individual tissue banks at Stanford to use STRIDE, in return they agree to make their biospecimen data discoverable. We provide a self-service tool that Stanford researchers use to determine if any of the participating banks contain specimens of interest. They can then contact the individual bank operators to request specimens.

Discussion

STRIDE is an effort to incrementally deploy a standards-based enterprise Informatics model supporting CTR, targeting three well described barriers: access to clinical data for research purposes, flexible research data management solutions and robust biospecimen data management and discovery. Implementing all three components within the same database model and system architecture offers a high level of data integration across components. In 2003, when the STRIDE project was initiated, the functionality of existing systems, such as caBIG and i2b2, particularly in relation to our need for a fully identified CDW, were not considered optimal. STRIDE is now a core part of Stanford's CTSA Informatics model. While there are currently no plans to implement STRIDE at sites other than Stanford, as the system is built using industry standard components, there is no reason why this could not be undertaken in the future.

References

1 Zerhouni EA. US biomedical research: Basic, translational, and clinical sciences. JAMA 2005, Sep 21;294(11):1352-8.

2 Sung NS, Crowley WF, Genel M, Salber P, Sandy L, Sherwood LM, et al. Central challenges facing the national clinical research enterprise. JAMA 2003;289(10):1278-87.

3 Payne PR, Johnson SB, Starren JB, Tilson HH, Dowdy D. Breaking the translational barriers: The value of integrating biomedical informatics and translational research. Journal of Investigative Medicine 2005;53(4):192-200

4 DiLaura R, Turisco F, McGrew C, Reel S, Glaser J, Crowley WF. Use of informatics and information technologies in the clinical research enterprise within US academic medical centers: Progress and challenges from 2005 to 2007. J Investig Med 2008, Jun;56(5): 770-9.

5 Williams RL, Johnson SB, Greene SM, Larson EB, Green LA, Morris A, et al. Signposts along the NIH roadmap for reengineering clinical research: Lessons from the clinical research networks initiative. Arch Intern Med 2008, Sep 22;168(17):1919-25.

6 Prokosch HU, Ganslandt T. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. Methods Inf Med 2009;48(1):38-44.

7 McConnell P, Dash RC, Chilukuri R, Pietrobon R, Johnson K, Annechiarico R, Cuticchia AJ. The cancer translational research informatics platform. BMC Med Inform Decis Mak 2008;8:60.

8 Murphy SN, Mendis M, Hackett K, Kuttan R, Pan W, Phillips LC, et al. Architecture of the open-source clinical research chart from informatics for integrating biology and the bedside. AMIA Annu Symp Proc 2007:548-52.

9 Brandt CA, Morse R, Matthews K, Sun K, Deshpande AM, Gadagkar R, et al. Metadata-Driven creation of data marts from an eav-modeled clinical research database. Int J Med Inform 2002, Nov 12;65(3):225-41.

10 Lyman JA, Scully K, Tropello S, Boyd J, Dalton J, Pelletier S, Egyhazy C. Mapping from a clinical data warehouse to the HL7 reference information model. AMIA Annu Symp Proc 2003:920.

11 http://www.oracle-base.com/articles/8i/VirtualPrivateDatabases.php

12 http://www.oracle.com/technology/products/text/index.html

13 Murphy SN, Barnett GO, Chueh HC. Visual query tool for finding patient cohorts from a clinical data warehouse of the partners healthcare system. Proc AMIA Symp 2000:1174.

14 Lin Z, Hewett M, Altman RB. Using binning to maintain confidentiality of medical data. Proc AMIA Symp 2002:454-8.

15 Scheufele EL, Dubey AK, Murphy SN. A study of the age attribute in a query tool for a clinical data warehouse. AMIA Annu Symp Proc 2008:1123.

16 Inmon WH. Building the Data Warehouse. Fourth Edition. Wiley Publishing 2005.

17 Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (redcap)-A metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2008, Sep 30.